Journal of Engineering Design and Analysis
Volume 3, Issue 2 - 2020, Pg. No. 65-71
Peer Reviewed Journal

Review Article

A Survey on Limitations of Traditional Outlier Detection Techniques for Wireless Sensor Networks

Biswaranjan Sarangi[1], Biswajit Tripathy[2]

[1]Synergy Institute of Technology, Bhubaneswar, Odisha, India.
[2]Gandhi Institute for Technological Advancement, Bhubaneswar, Odisha, India.

INFO

Corresponding Author:
Biswaranjan Sarangi, Synergy Institute of Technology, Bhubaneswar, Odisha, India.
E-mail Id:
biswaranjan.sarangi@gmail.com

How to cite this article:
Sarangi B, Tripathy B. A Survey on Limitations of Traditional Outlier Detection Techniques for Wireless Sensor Networks. *J Engr Desg Anal* 2020*;* 3(2): 65-71.

Date of Submission: 2020-11-22
Date of Acceptance: 2020-12-30

ABSTRACT

In Wireless Sensor Networks the major deviations from the sample of sensed data are regarded as outliers which include noise, errors and malevolent attack on the network. Detection of outlier is used to filter noisy data, locate faulty nodes, and find out interesting events. Conventional outlier detection techniques are not directly related to wireless sensor networks due to the nature of sensor data which is univariate and multivariate. This paper gives an outline of traditional outlier detection techniques and their limitations. Further we also discuss about a technique-based classification with characteristics of outlier data and their limitations.

Keywords: Outlier, Outlier Detection, WSN, Classification, Machine Learning

A Wireless Sensor Network (WSN) is basically the latest development of Moore's Law in the direction of the tininess and omnipresence of computing devices. Normally, a wireless sensor node composed of sensing, computing, communication, and power components. These components are integrated on a tiny single or multiple boards. Usually a wireless radio transceiver, a power source, a small microcontroller, and many type of sensors such as temperature, pressure, humidity, light, heat, sound, and vibration are out fitted with each node. A number of sensors used to monitor and collect information from the environment and send the information to a central location.

Wireless Sensor Networks (WSNs) have been used with success in critical application scenarios, such as remote patient health monitoring, environmental monitoring, structural monitoring of engineering structures and military surveillance, where the dependability of WSNs becomes an important factor. WSNs densely distributed over a geographically area and individual nodes autonomously communicate and interact with each other over the wireless medium. The information obtained from the WSNs has to be accurate and complete. Analysis of data collected from sensor at timely manner is of high importance. Raw data collected often suffer from inaccuracy and incompleteness due to following reasons: (i) the low cost sensor nodes having rigorous resource constraints such as battery power, computational capacity, memory and communication bandwidth; (ii) sensor nodes operations are frequently vulnerable to harsh and unattended environment effects and; (iii) sensor nodes are susceptible to malicious attacks such as dissent of eavesdropping, service attacks, and black hole attacks.[2] For making high data quality, more data reliability and effective, identification of erroneous data is essential.

In this paper we look at i) outliers in WSN, ii) desirable properties of outlier detection techniques, iii) compare the usefulness and limitations of the different techniques. The rest of paper is structured as follows: Section II describes

Journal of Engineering Design and Analysis (ISSN: 2582-5607)

the fundamentals of outlier and classification criteria of outlier detection techniques for WSN. Section III provides a technique-based classification to compare various detection techniques proposed in the literature. Concluding remarks given in last section.

Outlier

Maximum work in outlier detection originates from the field of statistics. At first Hawkins defined the term ''outlier'' as ''An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism''.[11] On the other hand, Barnett and Lewis defined it as ''An outlier is an observation or subset of observations that appears to be inconsistent with the rest of the set of data''.[10] In addition outliers can be defined as, "those measurements that significantly deviate from the normal pattern of sensed data".[21] Due to mechanical faults, changes in system behavior, deceptive behavior, human fault, instrument fault or merely through natural deviations in populations, outliers occur.

Outliers are of two types depending on the scope of data as local outliers and global outliers. If the density of an object in a data set deviates from the local area in which it occurs then it is a local outlier. This type of outlier exists when some of the observations at a sensor node are inconsistent comparing to the rest of data. The Local Outliers are also known as First Order Outliers. There are two variations one is each individual node identifies the inconsistent values depending on its historical values. Other is each sensor node collect readings of its neighboring nodes along with its historical readings to collaboratively identify the inconsistent values. The global outlier identification can be done at different levels in the network depending on the architecture of network. It uses the whole data set and considered as a special case of contextual outlier detection. Here the user can examine outlier in different context of interest.[14] The global outliers are also known as higher order outliers. All data are transmitted to the sink node in a centralized architecture to detect outlier. In aggregated architecture, data are collected within its controlling range through the aggregator from nodes to identify outliers. Individual nodes also can identify global outliers.

There are three outlier sources such as noise and error, events and malicious attacks. A noise-related measurement coming from a faulty sensor is an error. Outliers caused by errors may occur frequently, while outliers caused by events tend to have extremely smaller probability of occurrence.[3] An event is defined as a particular occurrence that changes the real-world state, e.g., forest fire, air pollution, etc. This kind of outliers usually lasts for a relatively long period of time and changes historical pattern of sensor data.

Outliers are measured in two scales: Scalar and Outlier score.[9] Each data measurement classify into normal or outlier class in scalar scale, which is a zero-one classification measure. The outlier score scale assign outlier score to each data measurements depending on the degree of measurement is considered as an outlier and provide a ranked list of outliers. We have to choose to analyze top n outliers having the largest outlier score which is fixed or we can use a cut-off threshold to select the outliers which is flexible.

In data dimension scenario sensor data can be viewed as data streams which are collected by sensor nodes. Data streams can be univariate or multivariate. The univariate streams are represented by a set of values read by a unique type of sensor where as multivariate streams are represented by a set of values coming from different sensors of the same sensor node. Sensor data tends to be correlated in both time and space.[12] Temporal correlation occurs at a single node location due to changes in data values over time. Spatial correlation occurs at a single node location due to comparison with neighboring nodes. Spatiotemporal correlation occurs through a number of node locations due to changes in data value over time and space.

The classification of Outlier detection techniques are stated as distance based and normal state model based which is a machine learning approach.

Classification

Distance based techniques use some distance measure notions from statistical distribution, nearest neighbor based or cluster based to detect outliers.

Statistical based: The statistical approach to outlier detection assumes a distribution or probability model for the given data set. Here a probability distribution model captures the distribution of the data and evaluate the data instances to well fit the model. If the probability of data instances is very low then it is declared as an outlier. Parametric or Non-parametric techniques are two classifications.

Parametric technique assumes accessibility of the knowledge about primary data distribution. Distribution parameters are then estimated from the available data. Based on type of distribution assumed, these techniques are further categorized as Gaussian- based and non-Gaussian based models i n Gaussian based model, usually distributed data are assumed. This technique depends on the spatial-temporal correlations of sensor data and uses two statistical tests to locally detect outliers.[16-18] It only deals with one dimensional outlier data. It requires too much memory for a node to store old values.

In Non-Gaussian based model the data are not normally distributed.[19] used a symmetric α stable (s α s) distribution which is not suitable for real sensor data to find out outliers in form of impulsive noise. This technique uses spatial-temporal correlation of sensor data to detect outliers locally. It reduces the communication cost due to local transmission. It also reduces computational cost as the cluster-heads carry out most of the computation tasks.

Hybrid outlier detection technique is a statistical technique which presents a semi-supervised, local outlier detection method to identify errors and detect events in ecological applications of WSN.

Non parametric techniques normally define distance measure notions between a new test instance and the statistical model. This technique does not assume availability of data distribution and use some kind of threshold to determine an outlier. A non- parametric method uses two kinds of approaches. First approach is based on estimating the density first and then used for classification. Second approach is based on choosing category directly.

Histogram based approaches: In this technique histogram data (one-dimensional) is collected and not the raw data to identify global outlier [20]. For centralized processing, here the sink uses histogram information to extract data distribution from the network and filters out the non-outliers. Recollecting more histogram information from the whole network causes high communication overhead which is the limitation in this approach.

Kernel based approaches: This technique approximates data distribution. It uses a kernel density estimator to identify outlier in streaming sensor data online.[21],[2] When the number of values in its neighborhood is smaller than a user specified threshold value, then the value is taken as an outlier. This technique does not enable to detect spatial outliers and the main problem is its high reliance on the defined threshold.

Nearest neighbor based approach: In data mining and machine learning, nearest neighbor classifiers are used. It analyzes a data instance by using several distance notions. It computes the distance between two data instances to measure the similarity. Euclidean distance is generally used as a distance measure. When a data instance is located far from its neighbor, it is declared as an outlier. Choice of appropriate input parameter is its limitations.

Clustering-based approach: Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. First the set is partitioned into groups based on data similarity using clustering and then labels are assigned to the relatively small number of groups. It is also called unsupervised learning. It is suitable for anomaly detection from temporal data. The testing phase for clustering based techniques is fast since the number of clusters against which every test instance needs to be compared is a small constant.[24] The choice of an appropriate parameter of cluster width is its limitation. Hierarchical and partitioning clustering is basically used.

Combined approach (nearest neighbor based and clustering based): A new clustering based approach combined with nearest neighbor based approach is proposed to classify outliers. This methodology consists of the following four steps.

- First the clustering algorithm is applied on all the sensory data to group data into clusters
- For each produced cluster, the outlier detection algorithm is applied to label each cluster as normal cluster or outlier cluster.
- This step is to classify the degree of outlier value (error or event).
- The last step is to compute the trustfulness of each sensor node to increase certainty in trusting a specific node.

Normal state model based: Machine learning technique are used for classification, regression and density estimation in various application areas such as speech recognition, spam detection, bio- informatics ,computer vision, fault detection and advertising networks. There are many reasons to apply machine learning in WSN. First it made WSN more adequate with dynamic environment and we could not make a mathematical model to describe their behavior.

Prediction based: Kalman filter is an estimation algorithm which estimates inaccurate hidden variables and provides a prediction of the future system state based on past estimations. We can use a kalman filter where we have uncertain data in dynamic system. The kalman filter is a recursive estimator and no history of observations is required.[6] Presents this approach for in-network outlier detection which utilizes spatial and temporal dependencies among sensor data. kalman filter provide good results due to optimality and convenient for online real time processing. It is having low computation, minimum storage capability and the accuracy of outlier detection is high.

Classification based: Classification is generally used in data mining and machine learning domain. Prior knowledge about the data set is not required in this technique. Classi fication based techniques learn a classi fication model using data set instances and later classify this to one of the training classes. This is called training phase. This technique is either supervised or unsupervised.

Support vector machine based: A SVM is a classiifcation method for both non-linear and linear data. A non-linear mapping is used which transform the original training data to a higher dimension. In this new dimension, linear optimal hyper plane is searched which separate the tuples from one class into another class. The SVM finds this hyper plane using support vector and margins. It is a binary classifier. It learns a classification model during the training phase and uses that model to classify any unseen or new arrived data. Generalization ability of classifier is to train classifier parameter well for partition of newly arrived or unobserved data.

Hyper plane, hyper sphere and quarter sphere based SVM uses Euclidian distance metrics and hyper ellipsoidal, centered ellipsoidal SVM uses mahalanobis distance based metrics to detect outlier. Hyper plane SVM has poor classification and generalization ability and not suitable for power restricted WSNs those positioned in harsh environment. The hyper sphere SVM performs good generalization ability but not feasible to implement on energy constrained WSNs as it suffers with quadratic optimization problem. The quarter sphere SVM has better classification performance than hyper plane and hyper sphere. This involves with linear optimization problem which reduces the computational complexity. Quarter sphere SVM technique considers spatial temporal correlation of sensor nodes, so they can handle both local and global outliers.[25] Hyper ellipsoid SVM suffers with quadratic optimization problem and more computational and memory usage cost than the hyper sphere SVM. Centered ellipsoid SVM approach considers both multivariate and streaming data, also consider spatial temporal correlations. Computational complexity, online outlier detection and event identification are limitations.

Bayesian network-based: This approach is based on probabilistic analysis which is categorized into three parts.

Naive-Bayesian Network model: Naïve Bayesian classifiers assume that the attribute value on a given class is not dependent on the values of other attributes. This assumption is known as class conditional independence which is made to simplify the computation and considered to be naïve. If the probability of a sensed reading in its class is smaller than that of being in other classes, it is considered as an outlier [13]. It is however does not specify how to decide a specific spatial neighborhood under the dynamic change of network topology. It deals with one-dimensional data only.

Bayesian Belief Network based: This network is a graphical representation of a probabilistic dependency model which consists of a set of interconnected nodes. Each node represents a variable in the dependency model. The causal relationships between these variables are represented by the connecting arcs. A belief network is defined by two components. The first is a directed acyclic graph, where each node represents a random variable, and each arc represents a probabilistic dependence. The second component defining a belief network consists of one conditional probability table for each variable. If the network structure is known and the variables are observable, then learning the network is straightforward.[12] This technique may not work well in presence of the dynamic network topology change.

Dynamic Bayesian Network based: Here the Bayesian networks represent the sequences of variables. Generalizations of Bayesian networks which represent and solve decision problems under uncertainty are known as influence diagrams. This technique uses DBNs to fast track changes in dynamic network topology of sensor networks. This technique identifies outliers by computing the posterior probability of the most recent data values in a sliding window. The data measurement that fall outside the expected value interval is considered as outliers.[26] This method can handle several data streams straight away.

Spectral decomposition based: The objective of this approach is to find normal modes of behavior in the data, by using principal components.

Principle Component Analysis: PCA is a technique which is used for analysis, by reducing multidimensional data sets to lower dimensions. PCA is mostly used as a tool in a exploratory data analysis and for making predictive models. PCA involves the calculation of the eigenvalue decomposition or singular value decomposition of a data set, usually after mean centering the data for each attribute. The results of PCA are considered in terms of component scores and loadings. PCA is computationally inexpensive, can be applied to ordered and unordered attributes, and can handle sparse data and skewed data. Multidimensional data of more than two dimensions can be handled by reducing the problem to two dimensions but in a lower dimensional space PCA finds the most correct data. The data is predictable in the direction of maximum variance and violation of this is considered as outliers.[14]

Active learning based: It is a supervised classiifcation approach which is based on probabilistic as well as nearest-neighbor approach. It does not need a training data in testing phase for which computation capacity is low.[27]

Density based: Identification of density based local outlier is a learning method which is based on density estimation that assumes the occurance of outlier density is far than the original data, and makes a link between neighborhood data and outlier data.[28]

Parzen window based: This method estimate the density from training set by using a non-parametric density estimation method to detect outlier.[28] It is a supervised method and can be worked in a semi-supervised manner. In semi- supervised method, the density is estimated from the typical instances with availablity of lebels. A threshold can be choosen to classify between outlier data and right one. The main problem is the communication cost in this method.

Challenges in Detection of Outliers in WSNs

- **High communication cost:** Minimization of communication overhead is a challenge to reduce the network traffic and extend network lifetime
- **Resource constraints:** Minimization of energy consumption is a challenge for the use of reasonable amount of memory required in storage and computation
- **Identifying outlier sources:** Difference between errors, events and malicious attacks as well as identification of outlier sources is a challenge
- **Distributed streaming data:** Due to dynamic change of distributed sensor data, online processing of distributed streaming data is a challenge
- Dynamic network topology and Communication
- **failures frequently:** Deployment of sensor network in unattended environment for a long period of time is vulnerable to dynamic network topology and regular communication failure
- Modeling normal objects and outliers effectively
- Application specific outlier detection

Table 1.Classification and Comparison of The Techniques with Limitations

Tech-niques	Data Dimension	Correlation	Outlier Type		Outlier Degree		Class	Limitations
			Local	Global	Scalar	Outlier score		
[16]	Univariate	Spatial	Collaboration			Fixed	Statistical based	Dynamic nature of WSN makes it difficult to select appropriate threshold value for evaluation. Non-parametric statistical model are not suitable for real time applications.
[17]	Univariate	Spatial & temporal	Collaboration			Fixed		
[18]	Univariate	Spatial & temporal	Collaboration			Fixed		
[19]	Univariate	Spatial & temporal	Individual	Aggregate		Flexible		
[20]	Univariate	Temporal		Centralized		Fixed		
[21]	Univariate	Temporal	Individual	Aggregate		Fixed		
[2]	Multivariate	Temporal		Aggregate		Fixed		
[22]	Univariate	Temporal				Fixed	NN based	Computational cost of handling multivariate data is more. Not scalable.
[23]	Univariate	Temporal		Aggregate		Fixed		
[15]	Univariate	Spatial & temporal		Aggregate		Fixed		
[24]	Multivariate	Temporal		Centralized		Flexible	Clus tering based	High computational complexity involved in measuring distance among multivariate data patterns
[25]	Multivariate	Temporal	Individual	Individual		Fixed	Classifi cation based	Some understanding about data distribution and correlation among sensor data is crucial. Not suitable for online outlier detection.
[13]	Univariate	Spatial & temporal	Collaboration		Scalar			
[12]	Multivariate	Spatial & temporal	Collaboration		Scalar			
[26]	Multivariate	Spatial & temporal	Collaboration		Scalar			

[14]	Multivariate	Spatial & temporal	Individual				Spectral Decom position based	Selecting suitable principal components which is needed to accurately estimate the correlation matrix of normal patterns is computationally very expensive

Conclusion

In this paper, we described the problems of outlier detection techniques in WSNs. We also provide information and desirable properties about outliers in WSNs. The shortcomings of existing techniques for WSNs clearly calls for developing outlier detection technique, which takes into account multivariate data and the dependencies of attributes of the sensor node, provides reliable neighborhood, proper and flexible decision threshold, and also meets special characteristics of WSNs such as node mobility, network topology change and making distinction between errors and events. Motivated by the need of outlier detection in wireless sensor networks, the shortcomings of the present techniques and keeping the research directions in view, it has been realized that there exists enough scope to improve the detection accuracy by using machine learning approach.

References

1. Zhang Y, Meratnia N, Havinga P. Outlier detection techniques for wireless sensor networks: a survey. *IEEE Commun Survey & Tutorials* 2010; 12(2): 159-170.
2. Subramaniam S, Palpanas T, Papadopoulos D et al. Online Outlier Detection in Sensor Data using Nonparametric Models. *J Very Large Data Bases* VLDB 2006.
3. Martincic F, Schwiebert L. Distributed Event Detection in sensor networks Proc. International Conference on Systems and Networks communications 2006: 43-48
4. Mohamed MS, Kavitha T. Outlier detection using support vector machine in wireless sensor network real time data. *Int J Soft Comput Eng* 2011; 1(2).
5. Hassan AF, Mokhtar H, Hegazy O. A heuristic approach for sensor network outlier detection. *Int J Res Rev Wireless Sens Networks* 2011; 1(4).
6. Shuai M, Xie K, Chen G et al. A kalman filter based approach for outlier detection in sensor networks", in Computer Science and Software Engineering. *International Conference* on 2008: 4: 154-157.
7. Box GJG, Reinsel G. Time Series Analysis: Forecasting & Control. 3rd Edition, Prentice-Hall, 1994.
8. Alsheikh M, Lin S, Niyato D et al. Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications. *IEEE Communication Surveys and Tutorials* 2015.
9. Chandola V, Banerjee A, Kumar V. Anomaly Detection: A Survey. Technical Report. *University of Minnesota* 2007.
10. Barnett, Lewis T. Outliers in Staistical Data, *New York*: John Wiley Sons 1994.
11. Hawkins DM. Identification of Outline, *London:* Chapman and Hall 1980.
12. Janakiram D, Mallikarjuna A, Reddy V et al. Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks, *Proc. IEEE* Comsware 2006.
13. Elnahrawy E, Nath B. Context-Aware Sensors, Proc. EWSN,2004.
14. Chatzigiannakis V, Papavassiliou S, Grammatikou M et al. Hierarchical Anomaly Detection in Distributed Large Scale Sensor Networks, Proc. ISCC, 2006.
15. Zhuang Y, Chen L. In-Network Outlier Cleaning for Data Collection in Sensor Networks, Proc. VLDB, 2006.
16. Wu W, Cheng X, Ding M et al. Localized Outlying and Boundary Data Detection in Sensor Networks. *IEEE Trans. Knowl. Data Eng* 2007; 19(8): 1145-1157.
17. Bettencourt LA, Hagberg A, Larkey L. Separating the Wheat from the Chaff: Practical Anomaly Detection Schemes in Ecological Applications of Distributed Sensor Networks, *Proc. IEEE International Conference on Distributed Computing in Sensor Systems* 2007.
18. Hida Y, Huang P, Nishtala R. Aggregation Query under Uncertainty in Sensor Networks, 2003.
19. Jun MC, Jeong H, Kuo CCJ. Distributed Spatio-Temporal Outlier Detection in Sensor Networks, Proc. SPIE, 2006.
20. Sheng B, Li Q, Mao W et al. Outlier Detection in Sensor Networks, Proc. Mobi Hoc 2007.
21. Palpanas T, Papadopoulos D, Kalogeraki V et al. Distributed Deviation Detection in Sensor Networks, ACM Special Interest Group on Management of Data 2003: 77-82.
22. Branch J, Szymanski B, Giannella C et al. In Network Outlier Detection in Wireless Sensor Networks, Proc. EEE ICDCS, 2006.
23. Zhang K, Shi S, Gao H et al. Unsupervised Outlier Detection in Sensor Networks using Aggregation Tree, Proc. ADMA,2007.

24. Rajasegarar S, Leckie C, Palaniswami M et al. Distributed Anomaly Detection in Wireless Sensor Networks. *Proc. IEEE ICCS* 2006.
25. Rajasegarar S, Leckie C, Palaniswami M et al. Quarter Sphere Based Distributed Anomaly Detection in Wireless Sensor Network. Proc. IEEE International Conference on Communications 2007; 3864-3869.
26. Hill DJ, Minsker BS, Amir E. Real-Time Bayesian Anomaly Detection for Environmental Sensor Data, Proc. 32nd Congress of the International Association of Hydraulic Engineering and research 2007.
27. Abe N, Zadrozny B, Langford J. Outlier Detection by Active Learning. Proceedings of the 12th ACM SIGKDD International Conference on Knowlege Discovery and Data Mining. Chicago II, USA 2013: 504-509.
28. Breunig MM, Kriegel HP, Ng RT et al. LOF: Identifiying Density based Local Outliers. SIGMOD Record 2000; 29: 93-104.